Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 75 (2015) 2677 – 2682

Energy Procedia

The 7th International Conference on Applied Energy – ICAE2015

Typical Day Detection for Long Term Price Forecasting

Alexey Raskin*, Petr Rudakov

National Research Nuclear University MEPhI, Kashirskoe sh.31, Moscow, 115409, Russia

Abstract

This paper presents a new method of grouping hours for electricity price long term forecasting. The main idea of new method is to detect hours and days with similar structure of generation based on supply curves comparison. It was tested on the Russia electricity market data and we got good results. We confirm the hypothesis that we can use supply curves as a characteristic of day or hour to define groups of hours with similar price dependency on consumption level. Proposed a method allows two times more accurately predict the price depending on consumption, based on the splitting all the days and hours into compact groups. After all we show how this technique can help with detection of outliers.

© 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of Applied Energy Innovation Institute

Clustering; Levenshtein distance; long-term forecast; price forecasting

1. Introduction

Long-term forecasting of electricity prices is a complex problem and depends on many factors. The consumption of electricity is one of the main factors affecting the price of electricity and if most of the factors (e.g. taxes or inflation) will affect the price of electricity is more or less uniformly, the same changes in consumption in different days or hours may have different impact on the price of electricity. Price in peak hours is more sensitive to the volume of consumption than at night for example.

For the problem of long term price forecasting surprisingly little work has been done. This problem is studied by [1,2,3]. But the main focus of these researches is on principal scheme of long-term forecasting or macroeconomics issues. Mathematical models and techniques are discussed rarely.

Our task was to create an algorithm that would allow grouping the hours and days in which dependencies of price on the volume of consumption are the same. It helps to make the forecast of prices depending on the volume of consumption significantly more precise. In other words, it will allow us to

* Corresponding author. Tel.: +7-926-379-0680;
E-mail address: a.a.raskin@gmail.com.

1876-6102 © 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of Applied Energy Innovation Institute
doi:10.1016/j.egypro.2015.07.663

specify the range of price fluctuations when changing consumption (for example, by increasing the use of energy-saving technologies that perhaps is not so important for Europe, but essentially concerns emerging markets, including Russia) more accurately.

Nomenclature

R	supply curve
R(i)	the i-th application in supply curve R
$P_R(i)$	the price of i-th application in supply curve R
$V_R(i)$	the volume of i-th application in supply curve R

2. The formalization of the problem

The scatter plot below shows the price versus the consumption for all hours of the first half of 2012. We can suggest that this data set consists of groups of hours in which price is linear function of consumption, but the coefficients of that functions are different for different groups. The aim of our research is to find such groups and offer a method of such clustering.

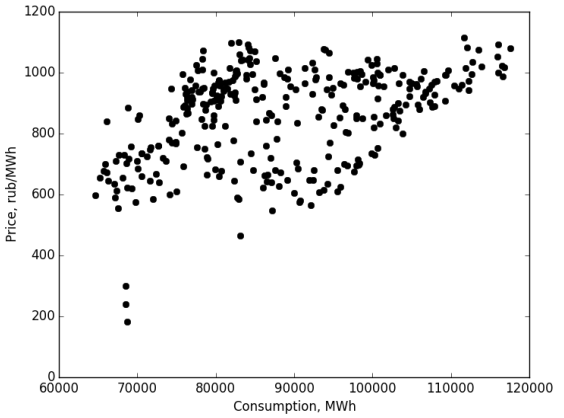


Fig. 1. Price versus consumption (Europe Zone of Russian Energy market, first half of 2012)

We should not only highlight the group in which the dependence of the price of the volume of consumption is strictly linear (there are a lot of ways to do that). Groups of points must be analyzable so that the analyst could give a formal description of these groups (or at least some of them).

One of the main characteristics of the state of the power system in a given hour is the supply curve. The intersection of the supply curve and the demand curve gives the price of electricity. We decide to form groups of similar hours and days to solve the problem described above based on comparison of supply curves. In this paper we will show that next proposition is true: similar in structure of the supply curve hours have similar linear regression coefficients (i.e., are on fig.1 on the same line).

The main goal was to make an algorithm for clustering all supply schedules into groups in which price of electricity will depend on consumption in linear way.

3. Data description

The dataset contains all curves in year of 2012 (8784 hours). There are about 197.9 applications in each hour in average so whole dataset contains 1 738 353 applications. All data is open and published on official site of market regulator (Administrator of Trading System). Some other energy markets published the same information (e.g. energy market Nord Pool Spot).

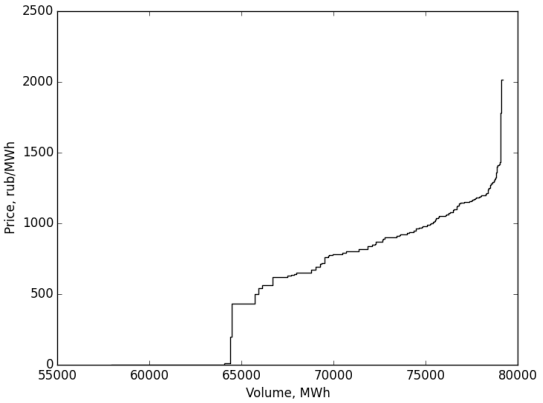


Fig. 2. Supply curve example

We split data into three groups: train data, validation set and test data. From train data we get clusters which define typical periods. On validation set we correct number of clusters and finally test our algorithm on test data. We get each group randomly in proportions 2:1:1.

4. Comparison of curves algorithm and clustering algorithm

Comparison between supply curves is not a typical problem. We need to develop a method of comparison of such data in order to use it in the process of clustering. Our aim was to define groups of similar objects using clustering algorithm k-means [4]. Since the supply curve is a sequence of applications of power stations we have turned our attention to the relevant group of proximity measures for sequences. We use Levenshtein distance [5] as a base of our distance. Levenshtein distance (or the editorial distance) shows a number of insertions, deletions and substitutions of elements necessary to make. However, Levenshtein distance cannot be directly used to comparison curves. We describe changes we made in it below.

$$\mathrm{Lev}_{\pi,\pi'}(\mathrm{i},\mathrm{j}) = \begin{cases} \max(\mathrm{i},\mathrm{j}) & \text{if } \min(\mathrm{i},\mathrm{j}) = 0 \\ \min \begin{cases} Lev_{\pi,\pi'}(i,j-1)+1 \\ Lev_{\pi,\pi'}(i,j-1)+1 \\ Lev_{\pi,\pi'}(i-1,j-1)+\left[\pi(i)\neq\pi'(j)\right] \end{cases} & ,else \end{cases}$$

Typical list of application is a sequence of pairs (price and volume) ordered from low to high price. So we modify Levenshtein distance to take into account not both parameters (price and volume). So let R and R' are two supply curves; $P_R(i)$ and $V_R(i)$ are price and volume of i-th application of R.

$$\mathrm{Lev}_{R,R'}(i,j) = \min \begin{cases} Lev_{R,R'}(i,j-1) + V_j \\ Lev_{R,R'}(i,j-1) + V_i \\ Lev_{R,R'}(i-1,j-1) + (V_j - V_i)\left[P(i) = P(j)\right] \end{cases}$$

We use distance function described above with k-medoids algorithm to find out groups of similar curves. This algorithm cannot detect number of clusters automatically so we make need make a number of clusterizations and choose the best one.

For defining number of clusters we use Dunn Index [6]. It shows relation between the maximum diameter of clusters and the minimum distance between clusters. Being defined in this way, the Dunn Index depends on the number of clusters in the set. If the number of clusters is not known apriori, the m for which the DI is the highest can be chosen as the number of clusters.

5. Results

We obtain three important results analyzing clustering results.

Hours which have similar curves (in conception of distance function we use) have similar coefficient of linear dependence between price and volume. It can be seen on fig. 3, where presented results of clustering: supply curves and pairs "price-volume". It is important result because it shows that distance function we use is meaningful.

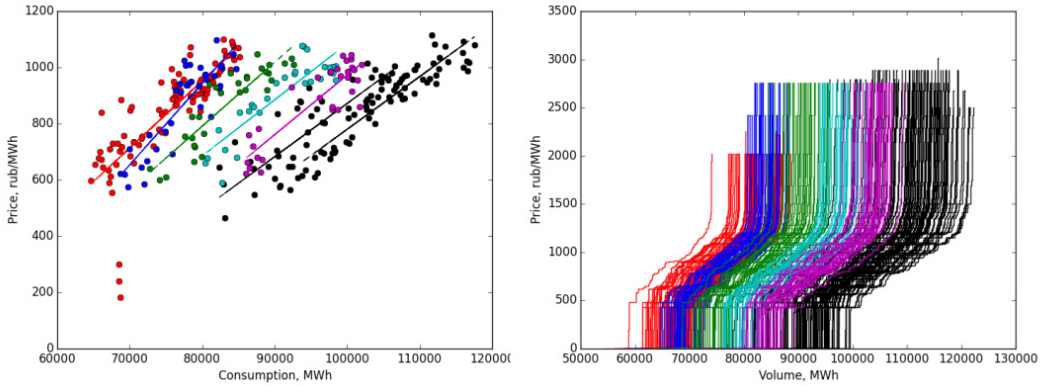


Fig. 3. Part of results of clustering (for easy visualization): price and consumption scatter plot (left) and curves (right)

The second important result we got is improving the quality of forecasting in comparison with using single approach. R-squared technique [7] was used for evaluating quality of forecasting. Linear regression models were created for each cluster we get (based on train data). Every hour from test data was assigned to nearest cluster and R-squared value was calculated for that point and corresponding linear model. On fig. 4 shown results of clustering and difference between one approximation line for whole data set and

after clustering. The R square error reduced twice after clusterization.

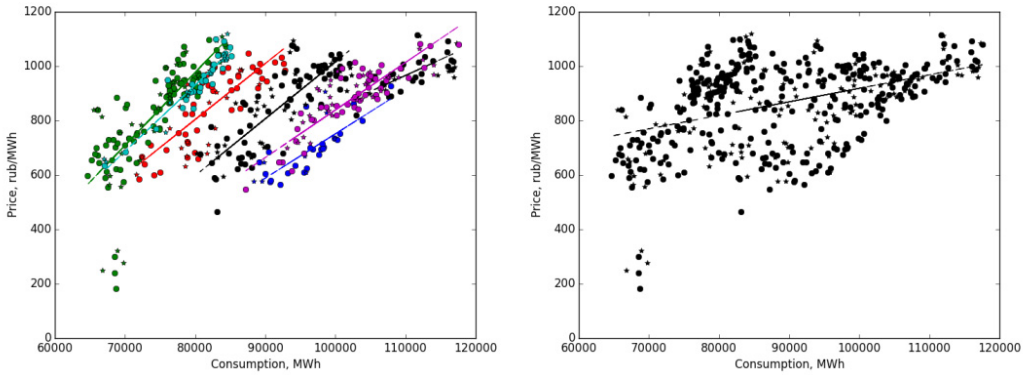


Fig. 4. Price and consumption scatter plot with approximation line for 7 clusters (left) and 1 cluster (right)

Lastly, presented approach can help us to define outliers (some days which has not normal structure of generation). On fig.3 we can see three points which staying alone. We cannot definitely say if these points are just days with low consumption (e.g. from cluster colored black) or outliers with something in structure of generation before clustering. After proposed procedure we can definitely say that these points are untypical points which are similar with “red” cluster.

The number of clusters is variable and depends on the purpose of analysis. On the proposed examples (fig.3), of a purely theoretical nature, it should be noted red and blue clusters. The point of these clusters are close, however, the dependence of the price of consumption is significantly different. The reason is that the blue cluster combines a peak hours (in the morning and evening), while red cluster contains mainly night and middle of the day.

6. Conclusion

In this paper was presented an approach of clustering days and hours into group with linear dependency price on consumption volume. It helps to reduce R-squared error of prediction in two times and find outliers in data set.

The most important part of investigation is a distance function which helps to compare supply curves and close link between compact group of curves and compact group of points on price and volume scatter plot. Hours with similar curves (where similarity is defined by presented approach) has strong linear dependency between price and electricity consumption.

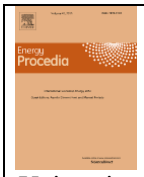
References

[1] Kononov, Yu.D. New requirements and approaches to the long-term forecasting of electricity prices. Power Tech, 2005 IEEE Russia, 2005

[2] Yanan Z, Li G, Zhou M, Lin S, Lo KL. An improved grey model for forecasting spot and long term electricity price. Power System Technology (POWERCON), 2010.

[3] Karsaz A, Mashhadi HR, Eshraghnia R. Cooperative co-evolutionary approach to electricity load and price forecasting in deregulated electricity markets. Power India Conference, 2006.

[4] Jain AK, Murty MN, Flynn PJ. Data Clustering: A Review. ACM computing serveys. 1999
[5] Levenshtein VI. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady, 1966
[6] Halkidi M, Batistakis Y, Vazirgiannis M. On Clustering Validation Techniques. J. Intell. Inf. Syst.; 17:2-3, p. 107-145, 2001
[7] Taylor JR. An introduction to error analysis. University Science Books, 1997



Biography

Alexey Raskin is senior analyst in consulting company in energetics. He has more than 8 years of experience in long term forecasting and strategy development. He analyzed more than 10 GWts of investment projects in Russia. Employee of the National Research Nuclear University “MEPhI” (Moscow, Russia).